Unsupervised Cross-lingual Representation Learning for Speech Recognition

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, Michael Auli

Facebook AI, USA
aconneau@fb.com, aabaevski@fb.com, locronan@fb.com, abdo@fb.com, michaelauli@fb.com

Abstract

This paper presents XLSR which learns cross-lingual speech representations by pretraining a single model from the raw waveform of speech in multiple languages. We build on wav2vec 2.0 which is trained by solving a contrastive task over masked latent speech representations and jointly learns a quantization of the latents shared across languages. The resulting model is fine-tuned on labeled data and experiments show that cross-lingual pretraining significantly outperforms monolingual pretraining. On the CommonVoice benchmark, XLSR shows a relative phoneme error rate reduction of 72% compared to the best known results. On BABEL, our approach improves word error rate by 16% relatively compared to a comparable system. Our approach enables a single multilingual speech recognition model which is competitive to strong individual models. We hope to catalyze research in low-resource speech understanding by releasing XLSR-53, a large model pretrained in 53 languages.

Index Terms: speech recognition, self-supervised learning, multilinguality, low-resource languages

1. Introduction

Cross-lingual learning aims to build models which leverage data from other languages to improve performance. This has been a long standing interest in the speech community [1, 2, 3, 4, 5, 6, 7] which includes systems able to transcribe multiple languages [8, 9, 10, 11, 12]. However, the vast majority of work in speech processing has focused on supervised cross-lingual training which requires labeled data in multiple languages. Unsupervised cross-lingual pretraining is often much scarcer than unlabeled speech and requires non-trivial human annotation.

Unsupervised representation learning, or pretraining, does not require labeled data and has received a lot of recent attention in computer vision [13, 14] after much success in natural language processing [15]. For the latter, cross-lingual pretraining has been shown to be very effective, particularly, for low resource languages [16, 17]. In speech processing, most work in this area has focused on monolingual unsupervised representation learning [18, 19, 20, 21, 22, 23, 24, 25].

In this paper, we focus on the cross-lingual setting by learning representations on unlabeled data that generalize across languages. We build on the pretraining approach of [26] which jointly learns contextualized speech representations as well as a discrete vocabulary of latent speech representations. The latter serves to effectively train the model with a contrastive loss (§2) and the discrete speech representations are shared across languages. Different to recent work on unsupervised cross-lingual pretraining, we fine-tune the Transformer part of the model instead of freezing all pretrained representations [27] or feeding them to a separate downstream model [28]. We extend the work of [27] by pretraining on multiple languages instead of just English and we experiment on top of a stronger baseline.

We evaluate XLSR on 5 languages of the BABEL benchmark [29] which is conversational telephone data and ten languages of CommonVoice [30], a corpus of read speech (§3). Multilingual pretraining outperforms monolingual pretraining in most cases, except for resource rich languages and we show that increased model capacity significantly closes the gap. We also demonstrate that XLSR representations can be fine-tuned simultaneously on multiple languages to obtain a multilingual speech recognition system whose performance is competitive to fine-tuning a separate model on each language (§4).

2. Approach

Unsupervised cross-lingual representation learning has shown great success by pretraining Transformers with multilingual masked language models [15, 16]. In this work, we learn cross-lingual speech representations by extending wav2vec 2.0 [26] to the cross-lingual setting. Our approach learns a single set of quantized latent speech representations which are shared across languages. Next, we outline the architecture (§2.1), training (§2.2) and adaptations for cross-lingual training.

2.1. Architecture

We follow the design choices described in [26]. The model contains a convolutional feature encoder \( f : \mathcal{X} \rightarrow \mathcal{Z} \) to map raw audio \( \mathcal{X} \) to latent speech representations \( z_1, \ldots, z_T \) which are fed to a Transformer network \( g : \mathcal{Z} \rightarrow \mathcal{C} \) to output context representations \( c_1, \ldots, c_T \) [15, 22, 25]. For the purpose of training the model, feature encoder representations are discretized to \( q_1, \ldots, q_T \) with a quantization module \( \mathcal{Z} \rightarrow \mathcal{Q} \) to represent the targets in the self-supervised learning objective (§2.2). A Gumbel softmax enables choosing discrete codebook entries in a fully differentiable way [31]. Each \( z_t \) represents about 25ms of audio strided by 20ms, the context network architecture follows BERT [15] except for relative positional embeddings [25].

2.2. Training

The model is trained by solving a contrastive task over masked feature encoder outputs. For masking, we sample \( p = 0.065 \) of all time steps to be starting indices and mask the subsequent \( \mathcal{M} = 10 \) time steps. The objective requires identifying the true quantized latent \( q_t \) for a masked time-step within a set of \( K = 100 \) distractors \( Q_t \) sampled from other masked time steps: \(- \log \frac{\exp(sim(c_t, q_t))}{\sum_{q \neq q_t} \exp(sim(c_t, q))} \) where \( c_t \) is the output of the transformer, and \( sim(a, b) \) denotes cosine similarity.

This is augmented by a codebook diversity penalty to encourage the model to use all codebook entries [32]. We maximize the entropy of the averaged softmax distribution over the codebook entries for each group \( p_{\theta} \) across a batch of utterances: \( \frac{1}{N} \log \frac{1}{N} \sum_{y=1}^{N} H(p_{\theta}) = \frac{1}{N} \sum_{y=1}^{N} \sum_{v=1}^{V} p_{\theta,v} \log p_{\theta,v} \). To stabilize the feature encoder we apply an L2 penalty over the outputs of the feature encoder.

When pretraining on \( L \) languages, we form multilingual batches [15, 16] by sampling speech samples from a multino-
We train a multilingual model in ten languages and monolingual (XLSR-Monolingual) to multilingual (XLSR-10) pretrained models (Base) fine-tuned individually on each language (ft=1). On CommonVoice, XLSR-10 obtains 13.6 PER on average (Avg), a relative PER reduction of 46.6% compared to monolingual models. We evaluate cross-lingual and monolingual models on each pretraining set. For comparison with the English model, we train Base and Large multilingual models on 1350h of speech audio from the 10 evaluation languages plus 557h of English audio. We sample low-resource languages with $\alpha = 0.5$ and train a model with $\alpha = 1$ for comparison (unbalanced). For multilingual fine-tuning, we either separate or share phoneme vocabularies across languages. For BABEL, we train a monolingual model on each of the 14 languages, as well as a Base and Large multilingual model on a total of 650h of speech audio in ten languages. Since the amount of data in each language is more balanced than for CommonVoice, we use $\alpha = 1$. The same speech audio is used for pretraining and fine-tuning and we use separate character sets for multilingual fine-tuning. We train a large model on the 53 languages of MLS, CommonVoice and BABEL, which consists of 56 thousand hours of speech data. We call this model XLSR-53 and make it publicly available.

4. Results

In our experiments, we first show that our approach is very effective for learning generic cross-lingual representations in an unsupervised way. Pretraining a single model on multiple languages significantly outperforms the previous state of the art on CommonVoice, as well as our own monolingual models. Second, we demonstrate the positive impact of cross-lingual transfer on low-resource languages and provide a better understanding of the trade-off between high-resource and low-resource languages. Finally, by fine-tuning a multilingual model on many languages at once, we show that we can obtain a single model for all languages with strong performance.

4.1. Unsupervised cross-lingual representation learning

In what follows, we compare XLSR to several baselines and show that unsupervised cross-lingual representation learning is very effective. We provide a comprehensive analysis of the impact of different pretraining methods on automatic speech recognition in Table 1 and 2.

4.1.1. Multilingual outperforms mono pretraining and prior art

We first compare monolingual (XLSR-Monolingual) to multilingual (XLSR-10) pretrained models (Base) fine-tuned individually on each language (ft=1). On CommonVoice, XLSR-10 obtains 13.6 PER on average (Avg), a relative PER reduction of
Table 1: CommonVoice phoneme error rate (PER). Models are pretrained on either one language (pt = 1) or 10 languages (pt = 10); and fine-tuned on each language (ft = 1) or all languages (ft = 10). D indicates the pretraining data, LS for English LibriSpeech (100h or 360h), BBL for BABEL (1070h), CV for English CommonVoice (1350h), CV mo for monolingual (see number of pretraining hours per language) and CV all for multilingual (1350h). Languages can be high-resource (es, fr, it) or low-resource (e.g. ky, sv, tr, tt).

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of pretraining hours per language</th>
<th>Number of fine-tuning hours per language</th>
<th>es</th>
<th>fr</th>
<th>it</th>
<th>ky</th>
<th>nl</th>
<th>ru</th>
<th>sv</th>
<th>tr</th>
<th>tt</th>
<th>zh</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>16h</td>
<td>1h</td>
<td>1h</td>
<td>29h</td>
<td>1h</td>
<td>1h</td>
<td>1h</td>
<td>1h</td>
<td>1h</td>
<td>1h</td>
<td>1h</td>
</tr>
<tr>
<td>Baselines from previous work</td>
<td></td>
<td></td>
<td>793h</td>
<td>1h</td>
<td>100h</td>
<td>360h</td>
<td>10h</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>m-CPC† [27]</td>
<td>LS 10h</td>
<td></td>
<td>38.7</td>
<td>49.3</td>
<td>42.1</td>
<td>40.7</td>
<td>44.4</td>
<td>45.2</td>
<td>48.8</td>
<td>49.7</td>
<td>44.0</td>
<td>55.5</td>
<td>45.8</td>
</tr>
<tr>
<td>m-CPC† [27]</td>
<td>LS 10h</td>
<td></td>
<td>38.0</td>
<td>47.1</td>
<td>40.5</td>
<td>41.8</td>
<td>40.5</td>
<td>42.5</td>
<td>43.7</td>
<td>47.5</td>
<td>47.3</td>
<td>42.0</td>
<td>55.0</td>
</tr>
<tr>
<td>Fer et al.† [40]</td>
<td>BBL 10l</td>
<td></td>
<td>36.6</td>
<td>48.3</td>
<td>39.0</td>
<td>38.7</td>
<td>47.9</td>
<td>45.2</td>
<td>52.6</td>
<td>43.4</td>
<td>42.5</td>
<td>54.3</td>
<td>44.9</td>
</tr>
<tr>
<td>Our monolingual models</td>
<td></td>
<td></td>
<td>12.6</td>
<td>11.7</td>
<td>7.0</td>
<td>14.0</td>
<td>9.3</td>
<td>20.6</td>
<td>9.7</td>
<td>7.2</td>
<td>22.8</td>
<td>12.3</td>
<td></td>
</tr>
<tr>
<td>XLSR-English</td>
<td>CV mo</td>
<td></td>
<td>13.7</td>
<td>20.0</td>
<td>19.1</td>
<td>13.2</td>
<td>19.4</td>
<td>18.6</td>
<td>21.1</td>
<td>15.5</td>
<td>11.5</td>
<td>27.1</td>
<td>17.9</td>
</tr>
<tr>
<td>XLSR-Monolingual</td>
<td>CV mo</td>
<td></td>
<td>6.8</td>
<td>10.4</td>
<td>10.9</td>
<td>29.6</td>
<td>37.4</td>
<td>11.6</td>
<td>63.6</td>
<td>44.0</td>
<td>21.4</td>
<td>31.4</td>
<td>26.7</td>
</tr>
<tr>
<td>Our multilingual models</td>
<td></td>
<td></td>
<td>13.6</td>
<td>15.2</td>
<td>11.1</td>
<td>18.1</td>
<td>13.7</td>
<td>21.4</td>
<td>14.2</td>
<td>9.7</td>
<td>25.8</td>
<td>15.3</td>
<td></td>
</tr>
<tr>
<td>XLSR-10 (unbalanced)</td>
<td>CV all</td>
<td></td>
<td>9.7</td>
<td>13.6</td>
<td>15.2</td>
<td>11.1</td>
<td>18.1</td>
<td>13.7</td>
<td>21.4</td>
<td>14.2</td>
<td>9.7</td>
<td>25.8</td>
<td>15.3</td>
</tr>
<tr>
<td>XLSR-10</td>
<td>CV all</td>
<td></td>
<td>9.4</td>
<td>14.2</td>
<td>14.1</td>
<td>8.4</td>
<td>16.1</td>
<td>11.0</td>
<td>20.7</td>
<td>11.2</td>
<td>7.6</td>
<td>24.0</td>
<td>13.6</td>
</tr>
<tr>
<td>XLSR-10</td>
<td>CV all</td>
<td></td>
<td>10.0</td>
<td>13.8</td>
<td>14.0</td>
<td>8.8</td>
<td>16.5</td>
<td>11.6</td>
<td>21.4</td>
<td>12.0</td>
<td>8.7</td>
<td>24.5</td>
<td>14.1</td>
</tr>
<tr>
<td>Our multilingual models(Large)</td>
<td></td>
<td></td>
<td>14.2</td>
<td>16.1</td>
<td>13.9</td>
<td>9.8</td>
<td>21.0</td>
<td>10.4</td>
<td>7.6</td>
<td>22.3</td>
<td>12.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XLSR-10</td>
<td>CV all</td>
<td></td>
<td>7.9</td>
<td>12.6</td>
<td>11.7</td>
<td>7.0</td>
<td>14.0</td>
<td>9.3</td>
<td>20.6</td>
<td>9.7</td>
<td>7.2</td>
<td>22.8</td>
<td>12.3</td>
</tr>
<tr>
<td>XLSR-10</td>
<td>CV all</td>
<td></td>
<td>8.1</td>
<td>12.1</td>
<td>11.9</td>
<td>7.1</td>
<td>13.9</td>
<td>9.8</td>
<td>21.0</td>
<td>10.4</td>
<td>7.6</td>
<td>22.3</td>
<td>12.4</td>
</tr>
<tr>
<td>Our Large XLSR-53 model pretrained on 56k hours</td>
<td></td>
<td></td>
<td>10.4</td>
<td>10.9</td>
<td>29.6</td>
<td>37.4</td>
<td>11.6</td>
<td>63.6</td>
<td>44.0</td>
<td>21.4</td>
<td>31.4</td>
<td>26.7</td>
<td></td>
</tr>
</tbody>
</table>

49% compared to XLSR-Monolingual (Table 1). On BABEL, XLSR-10 improves over XLSR-Monolingual by 22% relative WER (Table 2) and by more over supervised training (Training from scratch). Pretraining on multiple languages results in cross-lingual transfer and better speech representations.

Compared to prior work, XLSR-10 Large reduces PER by 72% relative to m-CPC [27] on CommonVoice (Table 1). For BABEL, the most comparable work is [33] and in Table 3, XLSR-10 shows a relative word error reduction of 16% compared to their BLSTM-HMM baseline. In Georgian, we also see that our XLSR-53 model, which leverages additional unsupervised data from other datasets, obtains 31.1 which outperforms the best system by [41]. Our approach is also not tuned to the limit, for instance we do not consider Transformer language models but only a 4-gram KenLM.

4.1.2. Multilingual outperforms English-only pretraining

To isolate the impact of multilingual training versus simply training on more data, we pretrain an English-only CommonVoice model (XLSR-English) on the same amount of data as the multilingual model (1350h) and compare the two. Table 1 shows that on average, XLSR-English significantly improves over the monolingual models (average PER of 26.7 vs. 17.9 PER) but multilingual pretraining performs even better at 13.6 PER, a 24% relative PER reduction over XLSR-English. This shows that adding more training data is not the only reason for the improved accuracy: the similarity between the languages used in pretraining and fine-tuning also plays an important role. To enable few-shot learning with self-supervised learning on languages that are more low-resource, it is thus important to pretrain in multiple languages at once.

4.1.3. XLSR representations transfer well to unseen languages

To better assess the cross-lingual transfer of the learned representations, we evaluate the XLSR-10 BABEL model on four languages not seen during pretraining. We fine-tune this model on each language, and compare it to monolingual models pretrained specifically on these languages. Table 2 shows that a multilingual model not pretrained on any data from the four languages, still outperforms XLSR-Monolingual, reducing average CER from 29 to 22.8 which compares to results from previous work of 36.8 CER [6]. This further suggests that the learned representations capture generic features of the speech signal which transfer to many languages.

4.2. Understanding cross-lingual transfer learning

In this section, we examine several properties of unsupervised cross-lingual representation learning for speech recognition. We show that it is particularly effective on low-resource languages, then describe the transfer-interference trade-off which benefits low resource languages but hurts high resource languages. Finally, we show that adding capacity is important for multilingual pretraining.

4.2.1. Cross-lingual transfer learning improves low-resource language understanding

Unsupervised cross-lingual representation learning and cross-lingual transfer are particularly effective on low-resource languages. On CommonVoice, the separation between high and low-resource languages is more salient than for BABEL. We distinguish between low-resource and high-resource based on the amount of available unlabeled speech data. For example, French and Spanish have 353h and 168h and are thus high-resource, while Swedish and Turkish have 3h and 11h and are low-resource.
Table 2: BABEL results on out-of-pretraining languages (CER). XLSR-10 provides strong representations for languages not seen during pretraining, outperforming monolingual models pretrained specifically on these languages.

<table>
<thead>
<tr>
<th>Model</th>
<th>#pt</th>
<th>#ft</th>
<th>as</th>
<th>tl</th>
<th>sw</th>
<th>lo</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of pretraining hours</td>
<td>55h</td>
<td></td>
<td>76h</td>
<td>30h</td>
<td>59h</td>
<td>220h</td>
<td></td>
</tr>
<tr>
<td>Number of fine-tuning hours</td>
<td>55h</td>
<td></td>
<td>76h</td>
<td>30h</td>
<td>59h</td>
<td>220h</td>
<td></td>
</tr>
<tr>
<td>Baselines from previous work</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cho et al. [6] (Mono)</td>
<td>10</td>
<td>1</td>
<td>45.6</td>
<td>43.1</td>
<td>33.1</td>
<td>42.1</td>
<td>41.0</td>
</tr>
<tr>
<td>Cho et al. [6] (Stage-2)</td>
<td>10</td>
<td>1</td>
<td>41.3</td>
<td>37.9</td>
<td>29.1</td>
<td>38.7</td>
<td>36.8</td>
</tr>
<tr>
<td>Our monolingual models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training from scratch</td>
<td>1</td>
<td>1</td>
<td>50.2</td>
<td>41.7</td>
<td>40.8</td>
<td>43.5</td>
<td>44.1</td>
</tr>
<tr>
<td>XLSR-Monolingual</td>
<td>1</td>
<td>1</td>
<td>34.8</td>
<td>25.4</td>
<td>26.8</td>
<td>29.1</td>
<td>29.0</td>
</tr>
<tr>
<td>Our multilingual models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XLSR-10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#ft=1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XLSR-10 (Large)</td>
<td>10</td>
<td>1</td>
<td>29.4</td>
<td>21.9</td>
<td>16.6</td>
<td>23.3</td>
<td>22.8</td>
</tr>
<tr>
<td>XLSR-53 (Large)</td>
<td>53</td>
<td>1</td>
<td>17.9</td>
<td>13.1</td>
<td>21.3</td>
<td>22.4</td>
<td>18.7</td>
</tr>
</tbody>
</table>

Monolingual models perform poorly on low-resource languages but this is where cross-lingual transfer is most effective: XLSR-10 reduces PER over XLSR-Monolingual by a relative 67% on Swedish, 72% on Turkish, 72% on Kyrgyz, and 64% on Tatar.

On BABEL, results in Table 2 show that XLSR-10 outperforms XLSR-Monolingual on all languages. When evaluating on all BABEL languages, we observed that the biggest gains for XLSR-10 were obtained on the four lowest-resource languages: Georgian (ka), Kurmanji (ku), Tokpisin (tp) and Swahili (sw).

We also observe in Table 1 strong cross-domain transfer from the MLS data to the CommonVoice data which are both read-speech. The XLSR-53 increases performance particularly on languages that are shared between the two datasets. For example, in Spanish and French XLSR-53 outperforms XLSR-10 by 4.8 and 5.9% absolute PER. This model also performs well on other languages, including BABEL (see Table 2) on which it reduces word error rate by 2.3% WER over XLSR-10.

4.2.2. The transfer-interference trade-off: high vs. low-resource

Per language results on CommonVoice (Table 1) show transfer-interference trade-off [43]: for low-resource languages (e.g. ky, nl, sv, tr, tt), multilingual models outperform monolingual models because of positive transfer, however multilingual models perform worse on high-resource languages (es, fr, it), due to interference. Data from multiple languages enables better speech representations that transfer to low-resource languages but the model also needs to share its capacity across languages which degrades performance on high-resource languages.

For a given model capacity, the language sampling parameter $\alpha$ (see §2) controls this trade-off. Table 1 shows that training according to the true language distribution, XLSR-10 (unbalanced) using $\alpha = 1$, performs less well than XLSR-10, where more capacity is allocated to low-resource languages via $\alpha = 0.5$. The sole exception being French, the language with the most data. On average the unbalanced model obtains 15.3 PER while the balanced model obtains 13.6.

4.2.3. Increasing capacity for a multilingual pretrained model

The interference problem can be alleviated by adding more capacity to the multilingual model [43, 17]: the gap between multilingual models and monolingual models for high-resource languages can be reduced by increasing model capacity. In this work, we only study the impact of adding more capacity to the multilingual model, by training an XLSR-10 Large model. On CommonVoice, the Large model reduces PER by relative 9.6% compared to Base, reducing average PER from 13.6 to 12.3.

There are no gains on very low-resource languages like Swedish but significant gains on Spanish, French and Italian. On BABEL, average CER is reduced by a relative 6.8%. This shows that the multilingual model benefits from more capacity overall, and in particular for high-resource languages.

4.3. One model for all languages (multilingual fine-tuning)

When we fine-tune the pretrained model on each language individually, then we end up with a different model for each language. On the other hand, multilingual speech recognition aims to build a single model for all languages that performs as well or better than individual monolingual models. Next, we investigate fine-tuning a single model on the labeled data of all languages (ft=10) to obtain a single multilingual model instead of fine-tuning each language separately (ft=1). Training batches are constructed by sampling audio samples from multiple languages without upsampling.

For CommonVoice Table 1 shows that the Base model with monolingual fine-tuning of XLSR-10 obtains 13.6 average PER which compares to 14.1 PER for multilingual fine-tuning. When increasing model capacity (Large), multilingual fine-tuning is competitive to monolingual fine-tuning: 12.3 average PER (ft=1) vs. 12.4 average PER (ft=N). Increasing capacity is particularly important when fine-tuning on large amounts of supervised data from many languages. Multilingual fine-tuning performs competitively to monolingual fine-tuning and enables us to have a single model for many languages.

5. Conclusion

In this work, we investigated unsupervised cross-lingual speech representations learned from the raw waveform. We show that pretraining on data in many languages improves both over monolingual pretraining and prior work, with the largest improvements on low-resource languages. Fine-tuning the model on multiple languages at once enables a single multilingual speech recognition model competitive to individually fine-tuned models.
6. References


